УДК 519.7:004.855.5

Комбинаторная оценка вероятности переобучения на основе кластеризации и покрытий множества алгоритмов

А. И. Фрей

Московский физико-технический институт (государственный университет)

Завышенность теоретических оценок обобщающей способности алгоритмов классификации остаётся открытой проблемой уже более сорока лет, начиная с работ В. Н. Вапника и А. Я. Червоненкиса [1]. На практике наиболее перспективным выглядит комбинаторный подход [2], в рамках которого уже удалось добиться улучшения качества логических закономерностей [3]. Данная работа направлена на дальнейшее повышение точности комбинаторных оценок вероятности переобучения за счет учета сходства между алгоритмами с близкими векторами ошибок.

Рассмотрим задачу классификации. Пусть $\mathbb{X} = (x_1, \ldots, x_L)$ — генеральная выборка из L объектов, A — множество алгоритмов классификации, $I : A \times \mathbb{X} \to \{0,1\}$ — бинарная функция потерь. Для произвольной подвыборки $U \subset \mathbb{X}$ определим *число* и *частоту* ошибок алгоритма $a \in A$, соответственно, как $n(a,U) = \sum_{x_i \in U} I(a, x_i)$ и $\nu(a,U) = n(a,U)/|U|$. *Методом обучения* называют отображение вида $\mu : 2^A \times 2^{\mathbb{X}} \to \{0,1\}^L$. Метод обучения ставит в соответствие произвольному множеству алгоритмов A и обучающей выборке $X \subset \mathbb{X}$ некоторый алгоритм $\mu(A, X)$. В данной работе рассматривается метод *пессимистической минимизации эмпирического риска* (ПМЭР), действующий по правилу $\mu(A,X) \in Arg \max_{a \in A(X)} n(a, X)$, где $A(X) = \mathrm{Arg} \min_{a \in A} n(a, X)$, $\forall A, X$.

Пусть $[\mathbb{X}]^\ell$ — множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины k. Следуя [2], определим *вероятность переобучения* $Q_\epsilon(A, \mathbb{X})$ как долю разбиений $X \sqcup \bar{X}$, при которых переобученность $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$ алгорима $a = \mu(A, X)$ превышает заданный порог $\epsilon \in (0,1]$:

$$Q_\epsilon(A, \mathbb{X}) = P\left[\delta(\mu(A, X), X) \geq \epsilon\right], \qquad (1)$$

где $P = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell}$, а квадратные скобки действуют по правилу $[\textit{истина}] = 1, [\textit{ложь}] = 0.$

Введем на A отношение частичного порядка: $a < b$ означает, что $I(a,x) \le I(b,x), \forall x \in \mathbb{X}$ и $a \ne b$. Если $a < b$ и $\exists! x \in \mathbb{X}$ такой, что $a(x) \ne b(x)$, то будем говорить, что a *предшествует* b, и записывать $a \prec b$.

Теорема 1. *Пусть множество алгоритмов* A *представлено в виде разбиения на непересекающиеся подмножества* $A = A_1 \sqcup \ldots \sqcup A_t$, *такие что внутри каждого* A_i *алгоритмы допускают равное число ошибок на полной выборке. Пусть* μ *— ПМЭР. Для каждого* A_i *рассмотрим порождающее и запрещающее множества* X_i *и* $X_i^{'}$*:*

$$X_i = \bigcup_{a \in A_i} \{x \in \mathbb{X} : \exists b \in A : a \prec b,\ I(a,x) < I(b,x)\},$$
$$X_i^{'} = \bigcup_{a \in A_i} \{x \in \mathbb{X} : \exists b \in A : b < a,\ I(b,x) < I(a,x)\}.$$

Пусть, кроме этого, каждое подмножество вложено в объемлющее множество: $A_i \subset B_i$, $i = 1, \ldots, t$. *Тогда для вероятности переобучения выполнена следующая оценка:*

$$Q_\epsilon(A, \mathbb{X}) \le \sum_{i=1}^{t} P_i\ Q_{\epsilon_i}(B_i, \mathbb{Y}_i), \qquad (2)$$

где $P_i = C_{L_i}^{\ell_i} / C_L^{\ell}$ *— верхняя оценка на вероятность* $P[\mu(A, X) \in A_i]$, $\mathbb{Y}_i = \mathbb{X} \setminus X_i \setminus X_i^{'}$, *и введены следующие обозначения:* $\epsilon_i = \frac{L_i}{\ell_i k_i} \frac{\ell k}{L} \epsilon + \left(1 - \frac{\ell L_i}{L \ell_i}\right) \frac{m_i}{k_i} - \frac{|X_i^{'}|}{k_i}$, $L_i = L - |X_i| - |X_i^{'}|$, $\ell_i = \ell - |X_i|$, $k_i = k - |X_i^{'}|$, m_i *— число ошибок алгоритмов из* A_i.

По результатам численого эксперимента оценка (2) оказывается точнее оценки, полученной методом порождающих и запрещающих множеств [3]. Кроме этого, новая оценка эффективно вычислима для семейств с существенно большим числом алгоритмов, т.к. сумма (2) содержит меньшее число слагаемых из-за кластеризации алгоритмов с близкими векторами ошибок.

Литература

1. *Vapnik V.N., Chervonenkis A.Y.* On the uniform convergence of relative frequencies of events to their probabilities. – Theory of Probability and Its Applications. – 1981. – N 16(2). – pp. 264-280.
2. *Воронцов К.В.* Точные оценки вероятности переобучения. – Доклады РАН. – 2009. – Т. 429. – №1. – С. 15-18.
3. *Vorontsov K.V., Ivahnenko A.A.* Tight combinatorial generalization bounds for threshold conjunction rules. – PReMI’11. – 2011. – pp. 66-73.